ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs

A. Weisheimer,[1] F. J. Doblas-Reyes,[1] T. N. Palmer,[1] A. Alessandri,[2] A. Arribas,[3] M. Déqué,[4] N. Keenlyside,[5] M. MacVean,[1,3] A. Navarra,[2] and P. Rogel[6]

Received 11 September 2009; accepted 13 October 2009; published 12 November 2009.

[1] A new 46-year hindcast dataset for seasonal-to-annual ensemble predictions has been created using a multi-model ensemble of 5 state-of-the-art coupled atmosphere-ocean circulation models. The multi-model outperforms any of the single-models in forecasting tropical Pacific SSTs because of reduced RMS errors and enhanced ensemble dispersion at all lead-times. Systematic errors are considerably reduced over the previous generation (DEMETER). Probabilistic skill scores show higher skill for the new multi-model ensemble than for DEMETER in the 4–6 month forecast range. However, substantially improved models would be required to achieve strongly statistical significant skill increases. The combination of ENSEMBLES and DEMETER into a grand multi-model ensemble does not improve the forecast skill further. Annual-range hindcasts show anomaly correlation skill of ~0.5 up to 14 months ahead. A wide range of output from the multi-model simulations is becoming publicly available and the international community is invited to explore the full scientific potential of these data. **Citation:** Weisheimer, A., F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel (2009), ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs, *Geophys. Res. Lett.*, *36*, L21711, doi:10.1029/2009GL040896.

1. Introduction

[2] Over the last years, multi-model ensembles (MMEs) have become powerful tools to account for uncertainties due to model error in dynamical model-based predictions on time scales from days to seasons and centuries. Their success in ensemble forecasting on seasonal time scales relies mainly on reducing an apparent overconfidence of all single-model ensembles, that is MMEs widen the ensemble spread while the average ensemble-mean error is reduced [*Weigel et al.*, 2008].

[3] Here, first results from a new MME for seasonal-to-annual predictions are presented based on five leading European global coupled climate models constructed as part of the ENSEMBLES project. The scientific basis for seasonal predictability lies in the slowly evolving components of the climate system, like the ocean or land surface, that act as boundary conditions for the atmosphere with its shorter intrinsic time scales. A prime example of a coupled atmospheric and oceanic phenomenon is the ENSO (El Niño/Southern Oscillation) event in the tropical Pacific, which is the dominant mode of seasonal and interannual climate variability. The ENSEMBLES MME forecast skill for tropical Pacific SSTs is demonstrated and compared with a previous-generation MME for seasonal forecasting (DEMETER [see *Palmer et al.*, 2004]).

[4] The scope of this paper is threefold: i) In Section 2, a documentation of the new ENSEMBLES MME and the set-up of the re-forecast experiments is given. ii) Results of an assessment of systematic errors and forecast quality in the tropical Pacific and progress over the DEMETER MME are presented in Sections 3 and 4. iii) A large number of atmospheric and oceanic data from the ENSEMBLES re-forecast experiments are becoming publicly available and strategies of data dissemination are described in Section 5. Section 6 summarizes and discusses the findings.

2. ENSEMBLES Seasonal-to-Annual Multi-model Experiments

[5] The ENSEMBLES MME for seasonal-to-annual forecasts comprises global coupled atmosphere-ocean climate models from the UK Met Office (UKMO), Météo France (MF), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Leibniz Institute of Marine Sciences at Kiel University (IFM-GEOMAR) and the Euro-Mediterranean Centre for Climate Change (CMCC-INGV) in Bologna. All models include major radiative forcings. None of the coupled models has flux adjustments. The atmosphere and ocean were initialized using realistic estimates of their observed states and each model was run from an ensemble of nine initial conditions. Table 1 summarizes the main model components and their initialization strategies. Further details on the initial condition perturbations can be found in the auxiliary material.[7]

[6] Retrospective forecasts, or hindcasts, that emulate real-time seasonal forecast situations for the past, were performed in a coordinated experiment by the above-described models. The common hindcast period of the ENSEMBLES MME

[1]ECMWF, Reading, UK.
[2]CMCC, Bologna, Italy.
[3]Met Office, Exeter, UK.
[4]Météo France, Toulouse, France.
[5]Leibniz-Institut für Meereswissenschaften an der Universität Kiel (IFM-GEOMAR), Kiel, Germany.
[6]CERFACS, URA1875, Toulouse, France.

0094-8276/09/2009GL040896$05.00

[7]Auxiliary materials are available in the HTML. doi:10.1029/2009GL040896.

Table 1. Overview of Models Contributing to the New ENSEMBLES Multi-model Ensemble

Partner	Atmospheric Model and Resolution	Ocean Model and Resolution	Initialization		Additional Components and Comments	References
			Atmosphere and Land	Ocean[a]		
ECMWF	IFS CY31R1; T159/L62	HOPE; 0.3°–1.4°/L29	ERA-40/oper. analysis, atmospheric singular vectors	wind stress perturbations to generate ensemble of ocean reanalyses; SST perturbations at initial time	operational seasonal forecasting system S3	T. N. Stockdale et al. (ECMWF Seasonal Forecast System 3 and its prediction of SST, manuscript in preparation, 2009) and *Balmaseda et al.* [2008]
UKMO	HadGEM2-A; N96/L38	HadGEM2-O; 0.33°–1°/L20	ERA-40/oper. analysis, anomaly assimilation for soil moisture	wind stress perturbations to generate ensemble of ocean reanalyses; SST perturbations at initial time	fully interactive sea ice module	*Collins et al.* [2008]
MF	ARPEGE4.6; T63	OPA8.2; 2°/L31	ERA-40/oper. analysis	wind stress, SST and fresh water flux perturbations to generate ensemble of ocean reanalyses	GELATO sea ice model	*Daget et al.* [2009] and *Salas Mélia* [2002]
IFM-GEOMAR	ECHAM5; T63/L31	MPI-OM1; 1.5°/L40	initial condition permutations of three coupled climate simulations from 1950 to 2005 with SSTs restored to observations	initial condition permutations of three coupled climate simulations from 1950 to 2005 with SSTs restored to observations		*Keenlyside et al.* [2005] and *Jungclaus et al.* [2006]
CMCC-INGV	ECHAM5; T63/L19	OPA8.2; 2°/L31	AMIP-type simulations with forced SSTs	wind stress perturbations to generate ensemble of ocean reanalyses, SST perturbations at initial time	dynamical snow-sea ice model and land-surface model	A. Alessandri et al. (The INGV-CMCC Seasonal Prediction System: Improved ocean initial conditions, submitted to *Monthly Weather Review*, 2009) and P. Di Pietro and S. Masina (CMCC-INGV global re-analyses, manuscript in preparation, 2009)

[a]See also auxiliary material.

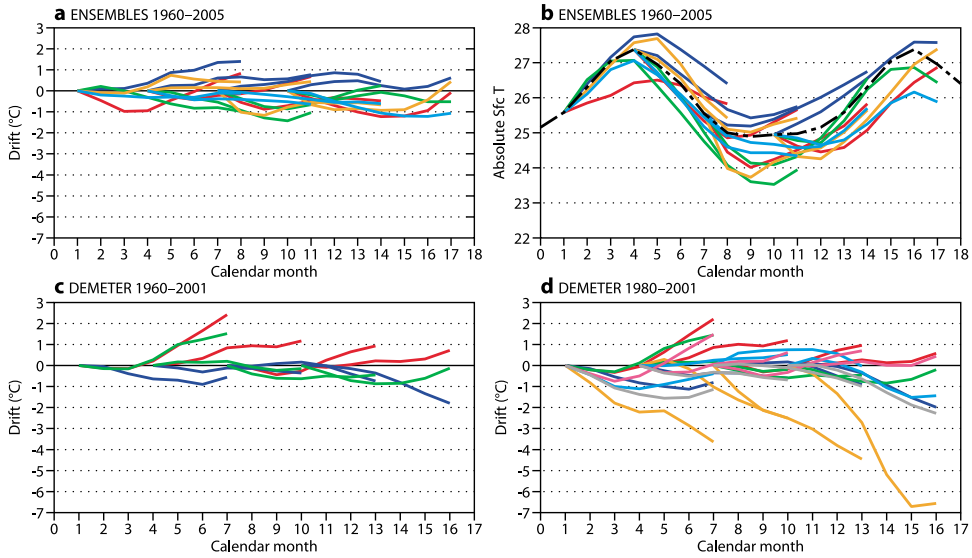

Figure 1. Systematic errors of the Niño3 SSTs in seasonal hindcasts. (a) Mean model drift relative to ERA-40 for the 5 ENSEMBLES models over the period 1960–2005, 4 start dates, up to 7 months lead-time; (b) as in Figure 1a but absolute SSTs. The black dashed line shows the annual cycle of the verification; c) as in Figure 1a but for the 3 DEMETER models over 1960–2001 and up to 6 month lead-time; d) as in Figure 1a but for the 7 DEMETER models over 1980–2001 and up to 6 months. Color code: red, MF; dark blue, ECMWF; green, UKMO; orange, IFM-GEOMAR; light blue, CMCC-INGV; pink, CERFACS; gray, LODYC.

covers the 46 years 1960–2005. For each year, 7-month-long seasonal forecasts starting on 1st of February, May, August, and November have been issued. Additionally, the November forecasts from all models except for CMCC-INGV were extended to 14-month-long annual forecast.

[7] The skill of the DEMETER MME for seasonal forecasts was computed for comparison. Since DEMETER, the models used in ENSEMBLES have improved in all aspects: in their physical parameterizations, by including additional components (e.g., sea-ice or land-surface modules) and interannual variability in the greenhouse gas forcing; in resolution and in the initialization. The DEMETER MME is available in two different configurations: a three-model ensemble over the hindcast period 1960–2001 and a seven-model ensemble covering the period 1980–2001. Similar to ENSEMBLES, the individual model ensembles consist of 9 ensemble members.

[8] There are several ways to construct MMEs by combining individual models [e.g., *Krishnamurti et al.*, 1999; *Doblas-Reyes et al.*, 2005]. However, given the relatively small sample size of seasonal hindcasts, finding robust non-equal weights in the combination of models proved difficult. Thus, the simplest and most straightforward approach by applying equal weights to all contributing models and ensemble members is used here.

3. Systematic SST Errors

[9] Although initialized using observations, seasonal forecast models develop, over the forecast time, systematic errors that lead the models to drift away from the observed state. Figure 1a shows the mean model drift for sea surface temperature (SST), estimated from all ensemble members and hindcasts, in the Niño3 region (5°S–5°N, 150°W–90°W) of the ENSEMBLES models for each of the four start months. The annual cycle of the model SSTs is in a good agreement with observations (Figure 1b). For comparison, Figures 1c and 1d show the SST drift for the DEMETER models. It is clear from Figure 1 that considerable progress has been made since DEMETER in reducing the systematic SST errors, in particular on longer lead-times. While the SST drift in DEMETER (Figure 1d) varied between +2° C and −7° C for up to 6 months lead, the ENSEMBLES models have a much reduced drift with an overall amplitude of less than ±1.5° C. Global maps of SST biases at different lead times for 5 DEMETER models and their corresponding ENSEMBLES models are shown in Figures S2 and S3. As can be seen, most models have reduced SST biases in the whole Tropics, not only over the Pacific. However, Figures S2 and S3 also point out that there are still substantial areas, e.g., over the cold upwelling regions at the eastern boundaries of the oceans, where systematic errors are large and have, despite all efforts, not much improved in ENSEMBLES. It is impossible to isolate specific reasons for the improvements in the tropical Pacific because the coupled model systems have undergone a number of changes to their complexity, physics, resolution and initialization, as mentioned above. Experiments to test all these changes separately are not available.

4. Forecasting SST Anomalies in the Tropical Pacific

[10] The systematic errors discussed in Section 3 have been corrected for computing forecast anomalies by linearly removing the long-term mean over the hindcast period for a given start date and lead-time. The corrections were applied in cross-validation mode (by leaving one out) in order to emulate real-time forecast conditions as closely as possible. As an illustration of the forecast anomalies, Figure S4 shows time series over the hindcast period 1960 to 2005 of Niño3 SST anomalies in DJF as forecasted 2–4 and 5–7

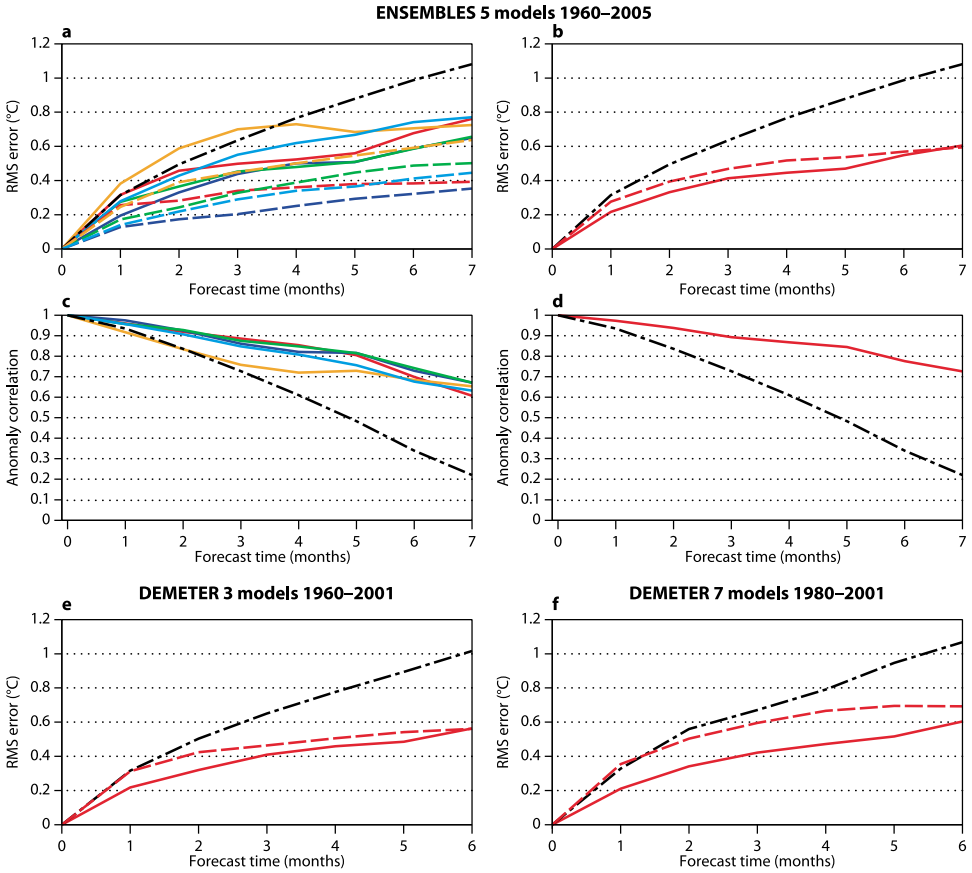

Figure 2. Niño3 SST RMSE (solid), ensemble standard deviation around the ensemble mean (dashed) and anomaly correlation (solid) as a function of forecast lead-time based on 4 start dates per year. (a–d) ENSEMBLES 1960–2005; (e) DEMETER multi-model over 1960–2001 and (f) DEMETER multi-model over 1980–2001. Multimodel results are shown in red and individual model ensembles Figures 2a and 2c in color as in Figure 1. The black dash-dotted curve indicates the performance of a persistence forecast.

months ahead. They show a very good agreement between the forecast and observed SST anomalies for all models and demonstrate how forecast uncertainty grows with forecast range. The correlations of the ensemble-mean of all models with the verification are highly statistically significant, even at longer lead times. The MME correlation is larger than, or equal to, the best correlation from the individual model ensemble.

[11] Figure 2 shows the temporal evolution of ensemble-mean root-mean square error (RMSE), ensemble spread and anomaly correlation for the Niño3 SST hindcasts anomalies. Figures 2a and 2c display all individual ENSEMBLES models, while Figures 2b and 2d show the MME. For comparison, a simple statistical persistence forecast is also given.

[12] In a perfect ensemble, over a large number of ensemble forecasts, the RMSE of the ensemble mean would equal the ensemble spread about the ensemble mean. A general feature of all single-model ensembles is, however, that the ensemble spread is substantially smaller than the RMSE (Figure 2a), that is, each individual ensemble is strongly underdispersive, or overconfident. As has been demonstrated in numerous studies [e.g., *Palmer et al.*, 2004; *Weigel et al.*, 2008], the multi-model combination effectively reduces the RMSE while the ensemble spread is increased leading to overall improved skill. For the ENSEMBLES MME SSTs this leads to an almost perfect match between the RMSE and spread (Figure 2b). Results for anomaly correlation in Figures 2c and 2d indicate that the MME also improves the correlation skill versus the individual models with correlations above 0.7 at 7-month lead-time.

[13] A similar evolution of RMSE and spread was found in the DEMETER three-model MME (Figure 2e). However, the seven-model DEMETER MME revealed, over the hindcast period 1980–2001, an overestimation of the ensemble spread (Figure 2f). Note that the MME evolution in Figures 2b and 2e would not change noticeably if computed over the 1980–2001 period.

[14] As a measure of probabilistic forecast skill Figure 3 shows the Brier skill score (BSS) of Niño3 SSTs for the ENSEMBLES, DEMETER and the combined grand ENSEMBLES & DEMETER MMEs. Where BSS = 1 the forecasts are perfect; BSS = 0 means the forecasts have as much skill as the reference, and BSS < 0 means less skillful than the reference forecast. The reference is the climatological forecast. The skill of the ENSEMBLES hindcasts is, on average, better than in DEMETER, especially for longer lead-times of 4–6 months (Figures 3c and 3d). Interestingly, this is a forecast range for which the systematic errors in ENSEMBLES are reduced the largest (cf. Figure 1). However, the uncertainty ranges of the BSS estimation as

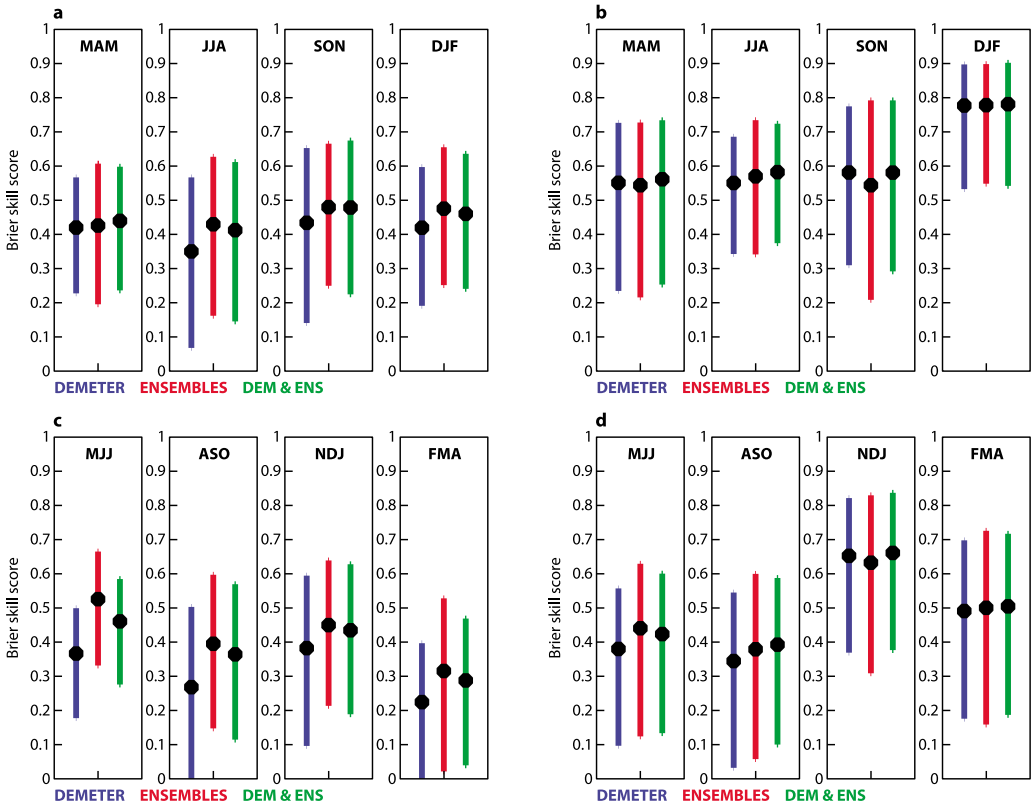

Figure 3. Brier Skill Score (BSS) for Niño3 SST hindcasts in DEMETER (blue), ENSEMBLES (red) and the combination of DEMETER and ENSEMBLES (green) with respect to a climatological forecast. The scores are for lead-times (a and b) 2–4 months and (c and d) 4–6 months and all 4 start dates over the period 1980–2001. The events considered are SST anomalies falling below the lower (Figures 3a and 3c) and above the upper (Figures 3b and 3d) terciles. The range of the bars indicates the 95% confidence interval estimates from 10.000 bootstrap re-samples with replacement.

expressed by the vertical bars in Figure 3, indicate that the differences are not significant.

[15] Hindcasts starting in November have been extended to 14 months in order to explore predictability on annual time scales. Figure 4 shows that there is some skill on these long lead-times for Niño3 SSTs. The anomaly correlation drops to 0.5 at month 9 and remains nearly constant thereafter. Remarkably, the above-mentioned good match between the RMSE and spread of the ensemble is further sustained over the extended forecast lead-time with an approximately linear error and spread growth.

5. Public Data Dissemination

[16] A common set of hindcast data from the ENSEMBLES models has been archived and is being publicly disseminated without charge for use in research, education and commercial work. The list of atmospheric variables contains daily and monthly mean data. The ocean output includes monthly means of ocean analyses and forecasts. Further details are given in the auxiliary material.

[17] Two dissemination systems, one based on the ECMWF Meteorological Archival and Retrieval System (MARS) and another one based on the Open-source Project for a Network Data Access Protocol (OPeNDAP), are provided to help users to access the ENSEMBLES data in the most efficient way for their specific requirements, see auxiliary material and http://www.ecmwf.int/research/EU_projects/ENSEMBLES/data/data_dissemination.html. The ENSEMBLES data are also available through the KNMI Climate Explorer, an interactive tool to analyze climate data [*van Oldenborgh and Burgers*, 2005].

6. Discussion and Conclusions

[18] A new re-forecast dataset for seasonal-to-annual time scales has been introduced based on an MME of five state-of-the-art coupled atmosphere-ocean circulation models. The MME has a smaller Niño3 SST RMSE and a much improved spread-skill relationship at all lead-times compared to any of the contributing single-model ensembles. Progress over the DEMETER MME includes a notable reduction of systematic errors and improved probabilistic forecast skill scores, in particular for longer lead-times of 4–6 months. While the main conclusions of this study also hold for other regions in the tropical Pacific, the exact degree of these improvements depends on the region, season, forecast lead-time and event.

[19] The combination of ENSEMBLES and DEMETER into a grand MME does not improve the forecast skill from ENSEMBLES any further (Figure 3). Why is that? The central reason for enhanced skill in a MME is the reduction of the overconfidence of the single-model ensembles, i.e., transforming the largely under-dispersive forecasts from each single-model ensemble into a better-dispersed MME [*Weigel et al.*, 2008]. The spread-skill relationship in the ENSEMBLES MME is already close to perfect. Therefore,

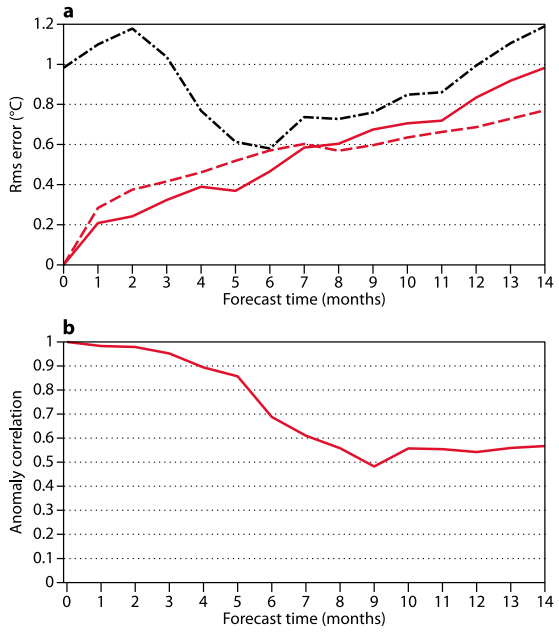

Figure 4. Niño3 SST annual-range (14 months) hindcasts of the ENSEMBLES multi-model over the period 1960–2005 for November start dates. (a) RMSE (red solid) and ensemble standard deviation (red dashed) as a function of forecast lead-time. The black dash-dotted curve indicates the skill of the climatological forecast. (b) Anomaly correlation. Note that CMCC-INGV did not contribute to the annual-range multi-model ensemble.

adding more ensemble members does not improve the spread-skill relation and related skill measures like the BSS. This effect is demonstrated in Figure S5 showing the RMSE and ensemble spread over lead-time for the grand DEMETER & ENSEMBLES MME. Similar to the DEMETER MME in Figure 2f, the grand MME is already over-dispersive (the same conclusion holds if the grand MME were to be constructed using only the 3-model DEMETER ensemble). In contrast, the ENSEMBLES MME in Figure 2b is well-dispersed as indicated by a better match between the RMSE and the ensemble spread. In such circumstances, the multi-model approach cannot further improve probabilistic forecast skill.

[20] For detailed discussions of more specific aspects of forecast quality the reader is referred to a follow-up paper (A. Alessandri et al., Evaluation of probabilistic quality and value of the ENSEMBLES multi-model seasonal forecasts: Comparison with DEMETER, submitted to *Geophysical Research Letters*, 2009).

[21] On the other hand, a statistically significant increase in forecast skill in the tropical Pacific over DEMETER is not achieved either by the ENSEMBLES MME or by the combined ENSEMBLES & DEMETER MME. This suggests that, to make substantial progress over the status quo in forecasting tropical Pacific SSTs fundamentally improved models that lead to intrinsically better probabilistic forecast skill are required. It is possible that this can only be achieved with substantially higher resolution models than are currently available [*Shukla et al.*, 2009].

[22] Extending the length of the hindcasts beyond seasonal time scales up to 14 months indicates an approximately linear growth of the RMSE and ensemble spread. No further degradation of the SST anomaly correlation after month 9 with $r \approx 0.5$ was found. Whether this will translate into useful forecast skill for applications remains to be demonstrated.

[23] A wide range of atmospheric and oceanic output from the ENSEMBLES MME simulations including ocean reanalyses are becoming publicly available and the international community is invited to explore the full scientific potential of these data.

[24] **Acknowledgments.** This work would not have been possible without the generous support from the modeling centers' staff. We especially thank A. Borrelli, M. Balmaseda, T. Stockdale, K. Mogensen, T. Jung, L. Ferranti, F. Molteni, C. Valiente, M. Fuentes, F. Venuti and J.-P. Piedelievre for their help and inspiring discussions. The comments of the reviewers helped to improve the manuscript. We acknowledge the ENSEMBLES project, funded by the European Commission's 6th Framework Programme through contract GOCE-CT-2003-505539.

References

Balmaseda, M. A., A. Vidard, and D. L. T. Anderson (2008), The ECMWF ORA-S3 ocean analysis system, *Mon. Weather Rev.*, *136*, 3018–3034, doi:10.1175/2008MWR2433.1.

Collins, W. J., et al. (2008), Evaluation of the HadGEM2 model, *Tech. Note HCTN 74*, Met Off. Hadley Cent., Exeter, U. K.. (Available at http://www.metoffice.gov.uk/publications/HCTN/index.html)

Daget, N., A. T. Weaver, and M. A. Balmaseda (2009), Ensemble estimation of background-error variances in a three-dimensional variational data assimilation system for the global ocean, *Q. J. R. Meteorol. Soc.*, *135*, 1071–1094, doi:10.1002/qj.412.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting. Part II: Calibration and combination, *Tellus, Ser. A*, *57*, 234–252.

Jungclaus, J. H., N. S. Keenlyside, M. Botzet, H. Haak, J.-J. Luo, M. Latif, J. Marotzke, U. Mikolajewicz, and E. Roeckner (2006), Ocean circulation and tropical variability in the coupled model ECHAM5/MPI-OM, *J. Clim.*, *19*, 3952–3972, doi:10.1175/JCLI3827.1.

Keenlyside, N. S., M. Latif, M. Botzet, J. Jungclaus, and U. Schulzweida (2005), A coupled method for initializing El Niño Southern Oscillation forecasts using sea surface temperature, *Tellus, Ser. A*, *57*, 340–356.

Krishnamurti, T. N., et al. (1999), Improved weather and seasonal climate forecasts from multi-model superensemble, *Science*, *285*, 1548–1550, doi:10.1126/science.285.5433.1548.

Palmer, T. N., et al. (2004), Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), *Bull. Am. Meteorol. Soc.*, *85*, 853–872, doi:10.1175/BAMS-85-6-853.

Salas Mélia, D. (2002), A global coupled sea ice-ocean model, *Ocean Modell.*, *4*, 137–172, doi:10.1016/S1463-5003(01)00015-4.

Shukla, J., R. Hagedorn, B. Hoskins, J. Kinter, J. Marotzke, M. Miller, T. N. Palmer, and J. Slingo (2009), Revolution in climate prediction is both necessary and possible: A declaration at the World Modelling Summit for Climate Prediction, *Bull. Am. Meteorol. Soc.*, *90*, 175–178, doi:10.1175/2008BAMS2759.1.

van Oldenborgh, G. J., and G. Burgers (2005), Searching for decadal variations in ENSO precipitation teleconnections, *Geophys. Res. Lett.*, *32*, L15701, doi:10.1029/2005GL023110.

Weigel, A. P., M. A. Liniger, and C. Appenzeller (2008), Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, *Q. J. R. Meteorol. Soc.*, *134*, 241–260, doi:10.1002/qj.210.

A. Alessandri and A. Navarra, CMCC, Via Donato Creti 12, I-40128 Bologna, Italy.

A. Arribas and M. MacVean, Met Office, FitzRoy Rd., Exeter EX1 3PB, UK.

M. Déqué, Météo France, 42 Avenue Coriolis, F-31057 Toulouse CEDEX, France.

F. J. Doblas-Reyes, T. N. Palmer, and A. Weisheimer, ECMWF, Shinfield Park, Reading RG2 9AX, UK. (antje.weisheimer@ecmwf.int)

N. Keenlyside, Leibniz-Institut für Meereswissenschaften an der Universität Kiel (IFM-GEOMAR), Duesternbrooker Weg 20, D-24105 Kiel, Germany.

P. Rogel, CERFACS, URA1875, 42 Avenue Coriolis, F-31057 Toulouse CEDEX 1, France.